

More Than *Obvious*: Better Methods *for* Interpreting NONDETECT DATA

Low-level concentrations of organic and inorganic chemicals are often reported as “nondetect” or “less-than” values. These concentrations are known only to be somewhere between zero and the laboratory’s reporting level (RL). Fifteen years ago, *ES&T* published an article titled “Less Than Obvious” (1) that outlined methods for the statistical analysis of nondetect data. Since that time, the fields of survival analysis and reliability analysis (2,3) have continued to improve on methods for handling censored data—those observations reported only as being above or below a threshold value. For example, methods for hypothesis tests and regression for censored data are now standard features of statistics software. Decades-old methods in these two fields that were originally applied only to “greater thans” can also be applied to the “less thans” of low-level environmental concentrations.

Yet, regulatory guidance for the environmental community has generally not incorporated these procedures. This article provides an overview of methods currently available for interpreting data with nondetects. More detail can be found elsewhere (4, 5).

Computing descriptive statistics

Nondetects occur in many disciplines, including air quality, water quality, astronomy, pharmacology, and ecology. Long considered “second-class” data, nondetects have complicated the familiar computations of descriptive statistics, tests of differences among groups, and development of regression models. Within environmental sciences, the most com-

mon procedure continues to be substitution of some fraction of the RL for nondetects, even though 15 years ago this was known to be wrong (1). The result is inaccurate statistics, poor and misleading regression models, and incorrect decisions about whether to remediate. There are better ways.

Current environmental guidance recommends three methods for computing descriptive statistics of data with nondetects: substituting one-half (or another fraction) of the RL; the delta-lognormal method (D-LOG), which was originally known as Aitchison’s method; and Cohen’s method (6–12). However, all three methods are considered old technology that exhibit either bias or higher variability than other methods now available.

Numerous studies have found that substituting one-half of the RL is inferior to other methods. Helsel and Cohn stated that the method “represents a significant loss in information” compared to other, better methods (13). Singh and Nocerino reported that it produced “a biased estimate of mean with the highest variability” (14), and Lubin et al. showed that it “results in substantial bias unless the proportion of missing data is small, 10 percent or less” (15). Resource Conservation and Recovery Act (RCRA) guidances recommend substitution only when data sets contain <15% nondetects, in which case the method is “satisfactory” (8, 12). However, that judgment appears to be based only on opinion rather than on peer-reviewed science. The U.S. EPA’s 2004 Local Limits Development Guidance Appendices break from this pattern by not recommending substitution methods (16). Instead, this guidance recognizes that substitution results in a high bias when the mean or standard deviation is calculated and that performance worsens as the proportion of nondetects increases.

Substitution introduces more problems today than 15 years ago, because most data today have multiple RLs. Several factors cause multiple RLs, including levels that change over time, samples with different dilutions, interferences from other constituents, different data interpretations for samples sent to multiple laboratories, or variations in RLs because methods for setting them have changed.

**Fifteen years later,
these methods are slow
to be used.**

DENNIS R. HELSEL
U.S. GEOLOGIC AL SURV EY

Regardless of the cause, substituting a fraction of these changing limits for nondetects introduces a signal unrelated to the concentrations present in the samples. Instead, the signal represents the pattern of RLs used. In the end, false trends may be introduced—or real ones canceled out.

Aitchison first applied his D-LOG method to economic data for which zero was a plausible value (17). As proposed, the method models detected observations with a lognormal distribution, with the assumption that all nondetects equal zero. The only difference between D-LOG and a simple substitution of zeros for all nondetects is how the mean of detected values is computed. Gilliom and Helsel found that the performance of D-LOG was essentially identical to that of zero substitution (18). Both methods had high rates of errors. Yet, D-LOG is still recommended in some guidance documents, including the EPA's Guidance for Data Quality Assessment (9).

EPA's Technical Support Document for Water Quality-Based Toxics Control modifies D-LOG, although the name remains the same (6). Nondetects are assumed to fall at the RLs rather than at zero. This change produces the highest possible value for the overall mean yet underestimates the standard deviation. The modified method has the same primary flaw as substituting the RL for all nondetects: The values substituted introduce a signal arising from changing RLs rather than from concentrations in the samples. Therefore, the poor performance of substituting the RL described by Gilliom and Helsel (18) and subsequent authors applies to EPA's modified D-LOG procedure. Hinton evaluated the modified procedure directly and found that better procedures outperformed it (19).

Cohen's method is based on maximum likelihood estimation (MLE), which fits the best lognormal distribution to the data (20). MLE requires more computing power than was available to most scientists when it was developed in the late 1950s. So, Cohen produced a lookup table of approximate coefficients to decrease the mean and standard deviation of detected observations, in order to estimate the mean and standard deviation of the entire distribution (20). The coefficients are a function of the proportion of nondetects in the data set. Cohen's method assumes that data follow a normal distribution and is developed for a single censoring threshold or RL.

Both assumptions are important limitations to how the method is applied today. Few modern data sets have only one RL, so data must be re-censored at the highest level before the tables can be used. For example, with RLs of 1 and 10 units, all detected observations between 1 and 10 (and all nondetects) must be designated as <10 units before the tables can be used. This assumption causes information to be lost, introducing error. Today, the lognormal distribution is considered more realistic than the normal distribution for most environmental data. Cohen's method is often computed with the logarithms of data, and estimates of mean and standard deviation of logarithms are transformed back into original units. This approach introduces a bias for

data with <50 observations (13, 21).

Cohen's method is now totally unnecessary. Today, statistical software can easily handle multiple RLs and provide more accurate solutions to maximum likelihood equations.

Current methods

Modern MLE software, imputation, and the Kaplan-Meier method are three more accurate methods for computing statistics on data with nondetects. Each is now available in the survival analysis or reliability analysis sections of commercial statistics software.

MLE solves a "likelihood equation" to find the values for mean and standard deviation that are most likely to have produced both nondetect and detected data. To begin, the user must choose a specific shape for the data distribution, such as the lognormal. Both detected observations and the proportion of data falling below each RL are used to fit the curve. MLE does not work well for data sets with <50 detected values, where 1 or 2 outliers may throw off the estimation, or situations in which insufficient evidence exists for one to know whether the assumed distribution fits the data well (13, 14, 21).

Imputation methods fill in values for censored or missing observations without assigning them all the same value. The distribution of data, and perhaps other characteristics, must be specified. For example, regression on order statistics (ROS) is a simple imputation method that fills in nondetect data on the basis of a probability plot of detects (13, 21). Multiple RLs can be incorporated. MLEs of mean and standard deviation can also be used to impute missing values (22).

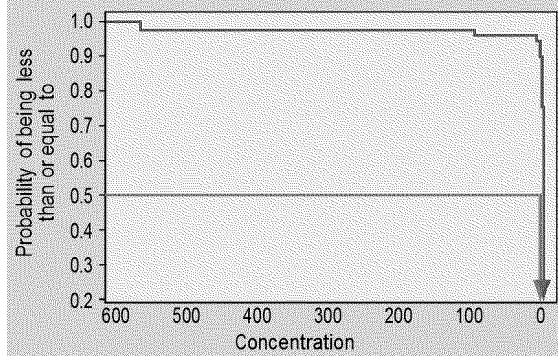
Because detected observations are used as measured, imputation methods depend less on assumptions of distributional shape than the MLE approach. As a result, imputation methods generally perform better than MLE with small sample sizes or when the data do not exactly fit the assumed distribution. For example, robust ROS estimates of mean and standard deviation performed better than MLE for sample sizes of <50 (13, 21). EPA (16) and the state of Colorado (23) have incorporated ROS methods into recent environmental guidance documents.

In medical and industrial statistics, Kaplan-Meier is the standard method for computing descriptive statistics of censored data (2, 3). It is a nonparametric method designed to incorporate data with multiple censoring levels and does not require specification of an assumed distribution. It estimates the percentiles, or cumulative distribution function (CDF), for the data set. The mean equals the area beneath the CDF (2). Kaplan-Meier is also a counting procedure. A percentile is assigned to each detected observation, starting at the largest detected value and working down the data set, on the basis of the number of detects and nondetects above and below each observation. Percentiles are not assigned to nondetects, but nondetects affect the percentiles calculated for detected observations. The survival curve, a step-function plot of the CDF, gives the shape of the data set (Figure 1).

FIGURE 1

Kaplan–Meier survival curve

This nonparametric method is designed to incorporate data with multiple censoring levels and to estimate the percentiles, or cumulative distribution function. The concentration scale goes from right to left because the data have been “flipped”. The red line is the data plot, and the blue line is the median value.



The Kaplan–Meier method has been used primarily for data with “greater thans”, such as time until a disease recurs. For this method to be applied to “less thans”, such as low-level chemical concentrations, data values must be individually subtracted from a large constant, or “flipped” (4), before the software is run. Flipping data is necessary only because of the way commercial software is now coded; it may become unnecessary with future versions as Kaplan–Meier becomes more widely used for analysis of “less-than” data. One caution is that estimates of the mean, but not percentiles, will be biased high with this method when the smallest value in the data set is a nondetect.

Testing hypotheses

Little guidance has been published for testing differences among groups of data with nondetects. The most frequently recommended method is the test of proportions, also called contingency tables (7, 8). This test is most appropriate for data with only one RL, because all the data will be placed into one of two categories: below or above the RL. Thus, the approach tests for differences in the proportion of detected versus nondetected data. Information is lost on the relative ordering between detected values; this is captured and used by nonparametric tests such as the rank-sum test. Moreover, the use of the test of proportions on data with multiple RLs requires that values must be re-censored and reported as either below or above the highest RL. Compared with methods that handle multiple limits, this approach loses information. Nevertheless, the primary advantages of the test of proportions are its simplicity and its availability in familiar software.

Unfortunately, the most commonly used test procedure is substituting one-half (or another fraction) of the RL before running standard tests such as the *t*-test. For data with one RL, Clarke demonstrated the significant errors produced by this procedure and by

imputation methods akin to the ROS method (24). Errors result from the arbitrary assignment of values to specific samples. The best results are obtained by first ranking the data (rankits) so that all nondetects are tied at the lowest rank (24). The subsequent *t*-test on the ranks approximates a nonparametric rank-sum test. As highlighted 15 years ago, standard nonparametric tests such as the rank-sum test work very well for analysis of data with one RL, whereas *t*-tests after substitution or imputation do not (1).

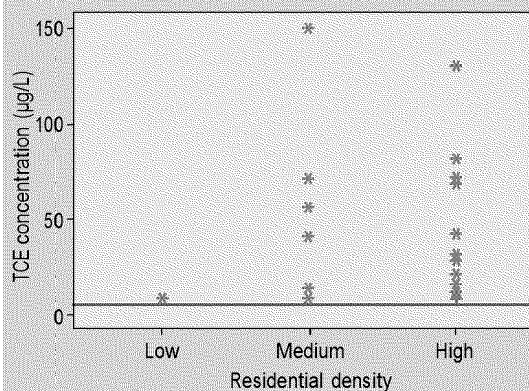
The Wilcoxon rank-sum and Kruskal–Wallis tests are sometimes recommended for comparing data with a single RL (7, 8). These nonparametric tests compare whether one group generally produces larger values than another. However, the Comprehensive Environmental Response, Compensation, and Liability Act (Superfund) guidance states that the Kruskal–Wallis test should not be used when >40% nondetects are present (7). Why this recommendation is made is unclear because no such limitations have been reported for these methods.

The RCRA guidance addendum makes the opposite recommendation: Use standard nonparametric tests rather than the test of proportions (8). Differences in the high ends of the distributions, if present, will be picked up by nonparametric tests, even at high levels of censoring. Groups will be found to differ if their proportions of nondetects differ, even if overall proportions are high (4). For example, statistically significant differences were found with the Kruskal–Wallis test (4) between the distributions of trichloroethylene (TCE) concentrations within the 3 groups of Figure 2, even though ~90% of the data are

FIGURE 2

Kruskal–Wallis test

Censored box plots of three residential densities with different patterns of trichloroethylene contamination, as determined by the Kruskal–Wallis test. Nondetects are 100% for low density, 91% for medium, and 80% for high. The reporting limit is 5 µg/L.



nondetects. Medium and high residential densities produced some high TCE concentrations, whereas the low-density residential group did not.

Guidance has been lacking on methods for testing data with multiple RLs. Both parametric and nonparametric methods for this situation were briefly

cited 15 years ago (1) and have not yet been adopted in environmental guidance documents. Now, however, much more detail is available (4). Parametric methods use MLE to perform tests equivalent to the *t*-test and analysis of variance (ANOVA) on data with multiple RLs. No substitution of fabricated values is required. Instead, likelihood-ratio tests determine whether splitting the data into groups explains a significant proportion of the overall variation. If so, the means differ among the groups.

Millard and Deverel pioneered the use of nonparametric score tests for environmental data in 1986 (25). These tests, sometimes called the generalized Wilcoxon or Peto-Prentice tests, extend the familiar Wilcoxon rank-sum and Kruskal-Wallis tests to data with multiple RLs. No values are substituted, and no re-censoring is necessary. The tests are used to compare the CDFs among groups of data and to determine whether their percentiles differ. Even if lower percentiles are indistinguishable because they are all nondetects, differences in higher percentiles will be seen if they are significant. The major impediment to the routine use of score tests has been commercial software that is coded to only recognize "greater thans", the form of censored data found in medical trials. Environmental data with "less thans" must first be flipped before current software can be used (4).

FIGURE 3

Survival function plot

Trichloroethylene concentrations in groundwater for low- (black), medium- (red), and high-density (green) residential areas were censored at 3 different RLs: 1, 2, and 5 µg/L. The generalized Wilcoxon test produces a *p*-value of 0.0003; this means that these 3 groups do not all have the same distribution and that differences exist between the upper ends of the curves. Adapted with permission from Reference 4.

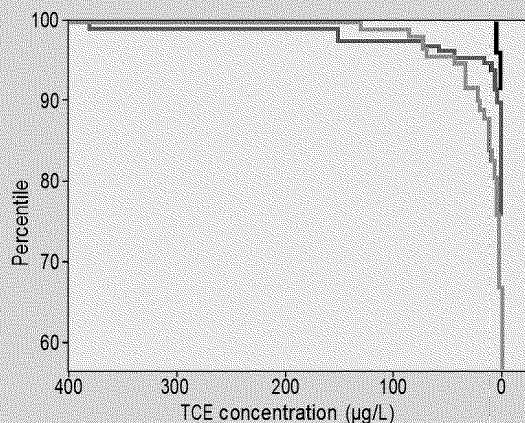


Figure 3 shows the percentiles of TCE concentrations for data in groups of low-, medium-, and high-density residential areas. The data are censored at 3 different RLs: 1, 2, and 5 µg/L. The generalized Wilcoxon test produces a *p*-value of 0.0003; this means that these 3 groups do not all have the same distribution. This is seen in Figure 3 as differences between the upper ends of the curves. The upper percentiles

of the low-density group remain low in concentration, whereas the medium- and high-density groups have higher concentrations. Even with 3 RLs and 90% censoring overall, the Wilcoxon test discerns that at least 1 group is different from the others. An MLE test on the logarithms of these data also finds significant differences.

Developing regression models

Regression equations are one of the foundations for interpreting environmental data. Fifteen years ago, Tobit regression was proposed for use with nondetect data (1). Now, more general methods for censored regression are readily available in commercial software but have not made their way into guidance documents or routine use by environmental professionals. Censored regression requires no substitution of fabricated data, so the pitfalls of that method can be avoided.

Slopes and intercepts for censored regression are fit by MLE rather than by least squares. This allows direct incorporation of nondetect data into model building. Likelihood ratio tests, rather than the familiar partial *t*- and *F*-tests, determine the significance of each explanatory variable. Likelihood statistics for models with and without that explanatory variable are compared to determine whether an explanatory variable belongs in the regression model. If these models fit the data equally, the *p*-value for that variable is high and the variable can be deleted from the model. Likelihood correlation coefficients are also available (4); analogs to most familiar regression statistics can be computed.

Imputation methods are also available for regression of censored data (15). MLE is used to fit slopes and intercepts on the basis of both censored and uncensored data. Values of the explanatory variables are then input to the regression model to impute values for the nondetect data. This method has been used to estimate values for concentrations of components, such as atrazine and its breakdown products (26). The imputed values are summed along with any detected values to estimate the total mass of herbicide. The quality of the imputed estimates depends on the fit of the regression model and on the amount of scatter around the regression line. Comparative testing of this versus other methods has not yet been done. However, if the regression equation is significant, then imputed values will outperform a simple substitution of one-half (or another fraction) of the RL for nondetects.

Nonparametric models for fitting straight lines to data with nondetects have advanced in the past 15 years. Lines based on Kendall's tau correlation coefficient have been applied to data in astronomy, in which light intensities often include "less-than" values (27). These nonparametric lines fit a median surface to data, rather than the mean surface of parametric regression. Outliers have much less influence on the Kendall-based lines. Another advantage of the Kendall procedure is that, unlike lines that use parametric MLE, an equation can be fit when both *x* and *y* variables are censored. Reference 4 provides examples of fitting these Kendall lines to environmental data.

Three opposing approaches

Three opposing approaches frequently emerge when the use of survival/reliability analysis methods for nondetects is discussed.

Substitute one-half (or another fraction) of the RL if only a few nondetects are present. The arguments in favor of this approach are that it is cheap and easy and that results can't be too far off when only a few values are substituted. On the other hand, it places values on these data that are more than is actually known, may introduce an artificial signal, and produces values based on arbitrary decisions because of the vagaries of how RLs are determined (anywhere from 2× to 10× the background standard deviation are common). Although MLE methods work poorly for small data sets, the situations for which this argument is most often made, imputation and Kaplan–Meier methods, work quite well without arbitrary substitutions.

Don't censor the data. Report the machine readings. The argument for this approach, although false, is that standard statistical methods could then be used. The argument against it is that RLs reflect the inability to determine whether observations differ from zero, or from one another. When the RL is 10, it cannot be stated with any confidence that machine readings of 2 and 4 are different. Declaring 4 to be larger than 2 in a hypothesis test is claiming more than is known. With data reported as "2 ± 10", weighting methods just as complex as survival analysis methods must be used to correctly perform computations.

Substitute the RL or delete nondetects in order to get a worst-case scenario. The argument for this approach is that a biased answer is better than none. However, excellent methods exist for getting an unbiased answer. A biased-high answer is seldom acceptable to the party who is paying for cleanup or prevention.

Until method precision increases to the point that RLs are not required, scientists must address the issue of handling nondetects. Given the importance, expense, and ramifications of environmental decision making, it is now more than obvious that environmental scientists should be using survival and reliability analysis methods to interpret data with nondetects.

Dennis R. Helsel is a geologist with the U.S. Geological Survey in Lakewood, Colo.

References

- Helsel, D. R. Less Than Obvious: Statistical Treatment of Data Below the Detection Limit. *Environ. Sci. Technol.* **1990**, 24, 1766–1774.
- Klein, J. P.; Moeschberger, M. L. *Survival Analysis: Techniques for Censored and Truncated Data*; Springer: New York, 2003.
- Meeker, W. O.; Escobar, L. A. *Statistical Methods for Reliability Data*; Wiley: New York, 1998.
- Helsel, D. R. *Nondetects and Data Analysis: Statistics for Censored Environmental Data*; Wiley: New York, 2005.
- Akritis, M. G. Statistical Analysis of Censored Environmental Data. In *Handbook of Statistics*; North Holland Publishing: Amsterdam, 1994; Vol. 12.
- Technical Support Document for Water Quality-Based Toxics Control; EPA/505/2-90-001; U.S. EPA, Office of Water: Washington, DC, 1991.
- Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites; EPA 540-R-01-003; U.S. EPA, Office of Emergency and Remedial Response: Washington, DC, 2002.
- Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities: Addendum to Interim Final Guidance; U.S. EPA, Office of Solid Waste: Washington, DC, 1992; www.epa.gov/epaoswer/hazwaste/cal/resoucel/guidance/sitechar/gwstats/gwstats.htm.
- Guidance for Data Quality Assessment: Practical Methods for Data Analysis; EPA/600/R-96/084; U.S. EPA, Office of Research and Development: Washington, DC, 1998.
- Assigning Values to Non-Detected/Non-Quantified Pesticide Residues in Human Health Food Exposure Assessments; Item 6047; U.S. EPA, Office of Pesticide Programs: Washington, DC, 2000; www.epa.gov/opppod01/trac/science/trac3b012_nonoptimized.pdf.
- Development Document for Proposed Effluent Limitations Guidelines and Standards for the Concentrated Aquatic Animal Production Industry Point Source Category; EPA-821-R-02-016; U.S. EPA, Office of Water: Washington, DC, 2002.
- RCRA Waste Sampling Draft Technical Guidance; EPA-530-D-02-002; U.S. EPA, Office of Solid Waste: Washington, DC, 2002.
- Helsel, D. R.; Cohn, T. Estimation of Descriptive Statistics for Multiply Censored Water Quality Data. *Water Resour. Res.* **1988**, 24, 1997–2004. (Note: In this paper, ROS is called MR.)
- Singh, A.; Nocerino, J. Robust Estimation of Mean and Variance Using Environmental Data Sets with Below Detection Limit Observations. *Chemom. Intell. Lab. Syst.* **2002**, 60, 69–86.
- Lubin, J. H.; et al. Epidemiologic Evaluation of Measurement Data in the Presence of Detection Limits. *Environ. Health Perspect.* **2004**, 112, 1691–1696.
- Local Limits Development Guidance Appendices; EPA 833-R-04-002B; U.S. EPA, Office of Wastewater Management: Washington, DC, 2004.
- Aitchison, J. On the Distribution of a Positive Random Variable Having a Discrete Probability Mass at the Origin. *J. Am. Stat. Assoc.* **1955**, 50, 901–908.
- Gilliom, R.; Helsel, D. R. Estimation of Distributional Parameters for Censored Trace Level Water Quality Data. 1. Estimation Techniques. *Water Resour. Res.* **1986**, 22, 135–146.
- Hinton, S. Delta Lognormal Statistical Methodology Performance. *Environ. Sci. Technol.* **1993**, 27, 2247–2249.
- Cohen, A. C. Simplified Estimators for the Normal Distribution When Samples Are Singly Censored or Truncated. *Technometrics* **1959**, 1, 217–237.
- Shumway, R. H.; et al. Statistical Approaches to Estimating Mean Water Quality Concentrations with Detection Limits. *Environ. Sci. Technol.* **2002**, 36, 3345–3353.
- Kroll, C. N.; Stedinger, J. Estimation of Moments and Quantiles Using Censored Data. *Water Resour. Res.* **1996**, 32, 1005–1012.
- Determination of the Requirement to Include Water Quality Standards-Based Limits in CDPS Permits Based on Reasonable Potential: Procedural Guidance; Colorado Water Quality Control Division: Denver, CO, 2003; www.cdphe.state.co.us/wq/PermitsUnit/wqcdpmt.html#RPGuide.
- Clarke, J. U. Estimation of Censored Data Methods to Allow Statistical Comparisons among Very Small Samples with Below Detection Limit Observations. *Environ. Sci. Technol.* **1998**, 32, 177–183.
- Millard, S.; Devereil, S. Nonparametric Statistical Methods for Comparing Two Sites Based on Data with Multiple Nondetect Limits. *Water Resour. Res.* **1988**, 24, 2087–2098.
- Liu, S.; et al. Analysis of Environmental Data with Censored Observations. *Environ. Sci. Technol.* **1997**, 31, 3358–3362.
- Akritis, M. G.; et al. The Theil–Sen Estimator with Doubly-Censored Data and Applications to Astronomy. *J. Am. Stat. Assoc.* **1995**, 90, 170–177.